The influence of speech rate on Fujisaki model parameters

Mixdorff, Hansjörg ; Leemann, Adrian ; Dellwo, Volker

Abstract: The current paper examines influences of speech rate on Fujisaki model parameters based on read speech from the BonnTempo-Corpus containing productions by 12 native speakers of German at five different intended tempo levels (very slow, slow, normal, fast, fastest possible). The normal condition was produced at an average rate of 6.34 syllables/s or 100%, the very slow version at 67%, and the fastest version at 161% of the normal rate. We extracted F0 contours and subjected them to decomposition using the Fujisaki model. We ordered all the data with respect to their actual speech rates. First, we assessed how prosodic realizations vary with speech rate and examined phrase command magnitudes, the number of phrase commands as well as the base frequency, accent command amplitudes, and the timing of accent command with respects to the underlying syllables and their nuclear vowels. Second, we analyzed between-sentence variability within and between speakers and investigated whether and how the prosodic structure is preserved at different speech rates. For very slow speech, we found for some of the speakers that the original phrase structure had disintegrated into something like a list of isolated words separated by pauses. Very fast speech became chains of uniform syllables at very high pitch and with almost flat intonation. With respect to the F0 range reflected by the amplitude of accent commands, we found strong interspeaker differences. While four of the subjects exhibited a significant reduction at higher speech rates, the others did not. As speed increases, it appears that F0 gestures commence earlier in the syllable, that is, the onset time of accent commands is located closer to the syllable/vowel onset than at lower speed.

DOI: https://doi.org/10.1186/s13636-014-0033-6

EURASIP Journal on Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

The influence of speech rate on Fujisaki model parameters

Hansjörg Mixdorff[1*], Adrian Leemann[2] and Volker Dellwo[2]

Abstract

The current paper examines influences of speech rate on Fujisaki model parameters based on read speech from the BonnTempo-Corpus containing productions by 12 native speakers of German at five different intended tempo levels (very slow, slow, normal, fast, fastest possible). The normal condition was produced at an average rate of 6.34 syllables/s or 100%, the very slow version at 67%, and the fastest version at 161% of the normal rate. We extracted F0 contours and subjected them to decomposition using the Fujisaki model. We ordered all the data with respect to their actual speech rates. First, we assessed how prosodic realizations vary with speech rate and examined phrase command magnitudes, the number of phrase commands as well as the base frequency, accent command amplitudes, and the timing of accent command with respects to the underlying syllables and their nuclear vowels. Second, we analyzed between-sentence variability within and between speakers and investigated whether and how the prosodic structure is preserved at different speech rates. For very slow speech, we found for some of the speakers that the original phrase structure had disintegrated into something like a list of isolated words separated by pauses. Very fast speech became chains of uniform syllables at very high pitch and with almost flat intonation. With respect to the F0 range reflected by the amplitude of accent commands, we found strong interspeaker differences. While four of the subjects exhibited a significant reduction at higher speech rates, the others did not. As speed increases, it appears that F0 gestures commence earlier in the syllable, that is, the onset time of accent commands is located closer to the syllable/vowel onset than at lower speed.

Keywords: Speech rate; F0 contour alignment; Fujisaki model; Prosodic structure

1 Introduction

To date, there are only relatively few accounts of the effects of speech rate on fundamental frequency F0. It is well established, for example, that an increase in speaking rate correlates with a decrease in pauses and a decrease in prosodic boundary marking (cf. [1,2]). Caspers and van Heuven [3] reported that rises in Dutch are steeper at fast articulation rates. Ladd et al. [4] showed that in accentual F0 rises, rise time becomes shorter the faster the articulation rate.

In the current paper, we will employ the well-known Fujisaki model [5] to examine the dependency of F0 contour on the syllable rate. This model reproduces a given F0 contour by superimposing three components: a speaker-individual base frequency Fb, a phrase component, and an accent component. The phrase component results from impulse responses to impulse-wise phrase commands associated with prosodic breaks. Phrase commands are described by their onset time T0, magnitude Ap, and time constant alpha. The accent component results from step-wise accent commands associated with accented syllables. Accent commands are described by on- and offset times T1 and T2, amplitude Aa, and time constant beta.

Within the framework of the Fujisaki model, Fujisaki and Hirose [6] found that phrase command magnitude Ap is lower for speakers with fast articulation rates than for speakers with a normal or slow rate. Fujisaki and Hirose [5] reported that accent command amplitudes Aa, too, decrease in faster articulation rate conditions. Moreover, they showed that faster speaking rate leads to the merging of accent commands. In his D.Eng. thesis [7], the first author already addressed the influence of the speech rate on the intonational features of German, however, on a very small data set. A single trained

* Correspondence: mixdorff@beuth-hochschule.de
[1]Department of Computer Science and Media, Beuth University Berlin, Luxemburger Str. 10, Berlin 13353, Germany
Full list of author information is available at the end of the article

speaker was asked to read a short text at comfortable (henceforth 'medium'), slow, and fast speeds. Analysis showed that the fast version was produced at a speed 28% higher and the slow version at a 15% lower speed than the medium one. Ap and Aa for the N↑ and I↓ intoneme, that is, accents with rising or falling F0, respectively (see Section 2), become smaller when speed increases. This means that the F0 range is reduced. However, for these parameters, the mean difference between fast and slow versions only amounts to 17% for an overall speed difference of about 50%. Interestingly, the change between slow and medium versions was greater than between medium and fast versions, though the difference in speed was not. As explained before, increased speech rate reduces the number of prosodic phrases and also reduces the duration of pauses. It also leads to the merging of some accent commands that are separate at lower speed. In more recent work on Swiss German, Leemann [8] showed that higher articulation rates can lead to a reduction of phrase boundaries, which has an effect on the other intonation phrases, making them overall longer in duration.

Although these results seem to indicate an inherent coupling between speech rate and the F0 contour, one also has to take into account that speakers employ individual strategies when producing speech at different velocities. In their well-known cineradiographic study, Kuehn and Moll [9] performed measurements of the velocity and displacement of the tongue during speech production and found considerable intra-speaker variation of these two parameters. It has also been shown that in fast speech, segment shortening tends to cause phonetic target undershoot or spatial reduction of articulatory targets [10], which leads to reduced articulatory displacement toward the phonetic goal and slower peak velocity of participating articulators [9]. More generally speaking, speaker-specific articulatory strategies are an important factor in explaining the articulatory variations [11]. Hence, it can also be expected that the impact of speech rate on F0 gestures will to certain extent be speaker-specific.

The remainder of this paper is structured as follows: Section 2 introduces the methodology for modeling German intonation adopted in this work. Section 3 discusses the corpus employed in this study and the prosodic features we extracted from it. Section 4 then presents results of individual samples as well as statistical analyses of the entire corpus. Section 5 concludes this paper offering a discussion of the findings and conclusions.

2 The concept of intonemes and their quantitative analysis

We will discuss some of the basics of the framework adopted in this study. In the works of Isačenko and Schädlich [12] and Stock and Zacharias [13], a given F0 contour is mainly described as a sequence of communicatively motivated tone switches, major transitions of the F0 contour aligned with accented syllables. Tone switches can be thought of as the phonetic realization of phonologically distinct intonational elements, the so-called intonemes. In the original formulation by Stock, depending on their communicative function, three classes of intonemes are distinguished, namely the N↑ intoneme ('non-terminal intoneme,' signaling incompleteness and continuation, rising tone switch), I↓ intoneme ('information intoneme' at declarative-final accents, falling tone switch, conveying information), and the C↑ intoneme ('contact intoneme' associated, for instance, with question-final accents, rising tone switch, establishing contact). Hence, intonemes in the original sense mainly distinguish sentence modality, although there exists a variant of the I↓ intoneme, I(E)↓ which denotes emphatic accentuation and occurs in contrastive, narrowly focused environments. Intonemes for reading style speech are predictable by applying a set of phonological rules to a string of text as to word accentability and accent group formation. Other F0 transitions - termed 'pitch interrupters' by Isačenko - will occur at phrase boundaries or in unstressed syllables where they do not have the same prominence-lending effect as tone switches (see [14]).

Based on this concept, Mixdorff and Jokisch [15] developed a model of German prosody anchoring prosodic features such as F0, duration, and intensity to the syllable as a basic unit of speech rhythm. In order to quantify the interval and timing of the tone switches with respect to the syllabic grid, the framework adopts the Fujisaki model for parameterizing F0 contours [1]. In a perception study [16] employing synthetic stimuli of identical wording but varying F0 contours created with the Fujisaki model, it was shown that information intonemes are characterized by an accent command ending before or early in the accented syllable, creating a falling contour. N↑ intonemes were connected with rising tone switches to the mid-range of the subject connected with an accent command beginning early in the accented syllable and plateau-like continuation up to the phrase boundary, whereas C↑ intonemes required F0 transitions to span a total interval of more than 10 semitones and generally starting later in the accented syllable, although the F0 interval was a more important factor than the precise alignment.

In the current study, we investigate the influence of speech rate on the realization of F0 contours. Tempo is a factor which we so far did not vary systematically within a larger range in our studies. We wish to explore how the three components of the Fujisaki model are influenced under different tempo conditions:

(1) On the utterance level: The base frequency Fb marks the floor of the F0 pattern. So far, we regard

Fb as a speaker-individual constant varying only slightly. However, we have observed that Fb can also vary depending on the emotional content of an utterance, for instance [17].

(2) On the phrase level: The phrase magnitude Ap reflects the degree of declination line reset at phrase boundaries. Earlier work suggested that Ap decreases as the tempo rises. We also expect to find fewer phrase commands at higher tempos as prosodic phrases will merge.

(3) On the syllable level: The accent command amplitude Aa corresponds to the interval of local F0 excursions associated with accented syllables and boundary tones. So far, we assume that Aa decreases with increasing tempo, and due to accent command merging or suppression of secondary accents, there will be fewer accent commands. The accent command onset times T1 and accent command offset times T2 with respect to the underlying syllable or nuclear vowel onset or offset times reflect the precise alignment of F0 excursions with the segmental tier. We hypothesize that increased speech rate also requires the F0 gesture to occur earlier in the syllable. We also wish to examine whether accent commands and hence the F0 gestures are more strongly anchored to the nuclear vowel onset than to the onset of the syllable.

3 Speech material and method of analysis

The speech material used in the current study are the recordings of the German L1 speech from the BonnTempo-Corpus [18,19]. It contains data from four male and eight female native speakers of standard German. The corpus is based on readings of a text from a novel by Schlink [20] of 76 syllables in three sentences (four main and three subordinate clauses). Versions at different tempos were elicited as follows: Subjects were provided the text and asked to familiarize with it by reading it aloud several times. Subsequently, they were recorded performing the task to read the text in a way they considered 'normal reading'. After that, subjects were recorded twice, the first time being instructed to read the text 'slowly' and the second time to read the text 'even slower'. In a third step, subjects were recorded under the instruction to read the text 'fast' and were then encouraged to read the text 'faster' until they considered themselves having reached a maximum reading speed or until their performance seriously deteriorated. From the resulting materials, five versions are examined in the current study: normal (no), slow (s1), even slower (s2), fast (f1), and fastest (f2). These were labeled on the syllabic level by the third author and his colleagues. In addition, they labeled the nuclear vowels. We are aware that the syllabic rate as a correlate of speech rate is inferior to the perceptual local speech rate (PLSR) proposed by Pfitzinger [21], as the local syllable rate and the local phone rate are not well-correlated, since they represent different perceptual aspects of speech rate. Perception experiments with short stretches of speech being judged on a rate scale revealed that neither syllable rate nor phone rate is sufficient to predict the perception results. Subsequently, it was shown that a linear combination of the two measures yielded a correlation of $r = 0.91$ and a mean deviation of 10% which is accurate enough to successfully extract PLSR from large spoken language corpora. However, the BonnTempo-Corpus does not contain manually corrected phone labels. Since the Fujisaki model commands are anchored to the syllabic layer (see Section 2) and we did not require an exact local estimate of speech rate, but a broad classification of speech rate on the utterance level, the following investigation is performed with respect to the syllabic rate. Based on the underlying text of the utterances, we marked all lexically stressed syllables of content words.

F0 values were extracted at a step of 10 ms with F0 floors and ceilings for male (50 to 300 Hz) and female participants (120 to 400 Hz) using the *PRAAT* default method [22]. All F0 contours were then subjected to Fujisaki model parameter extraction [23], with an alpha of 2.0/s, beta of 20.0/s and variable Fb. Results were checked and if necessary corrected in the *FujiParaEditor* [24]. Evaluating the alignment between phrase and accent commands with respect to the underlying syllables, while taking into account the status of these syllables as either being lexically stressed or not, we associated each accent command with the closest syllable. As explained in Section 1, a rising tone switch is invariably connected with the onset of an accent command and a falling tone switch with an offset of an accent command. All other accent command onsets or offsets are related to pitch interrupters at unaccented syllables.

Statistical analysis of tone switch alignment indicated that rising tone switches are most closely linked to syllable onsets, whereas falling tone switches are more closely aligned with syllable offsets. It has also been observed that falling or rising tone switches related to accented syllables do not necessarily occur during those syllables but before or after, respectively. Therefore, the search for the best alignment option has to include the neighboring syllables. Once the locations of stressed syllables have been scanned for accent commands nearby, the rest of the commands are aligned with the closest syllable based on a criterion of maximum overlap, Figure 1 shows the most important alignment options taken into account.

With respect to phrase command locations, the first command in an utterance will always be associated with the first syllable of that utterance. Due to the rise-fall

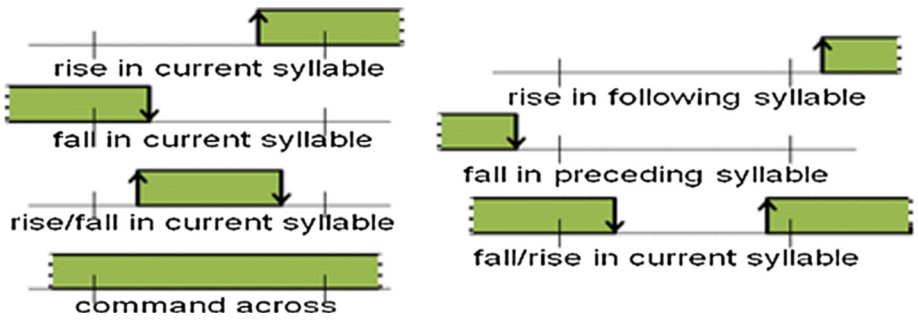

Figure 1 Most important alignment options for linking accent commands with underlying syllables.

characteristics of the phrase component, however, the phrase command usually occurs a few hundred milliseconds before utterance onset and the maximum of the phrase component ideally coincides with the segmental onset (see example in Figure 2). Subsequent phrase commands can be linked to following phrases, especially when these are preceded by short pauses. However, at many shallow boundaries, phrase commands will not appear as we will also see in the material examined in this study.

4 Results of analysis

4.1 General observations

Table 1 lists syllable rates for all speakers in all five different tempo conditions. Besides, syllable rates were expressed as percentages of the rate calculated for the normal condition for each speaker set to 100%. As can be seen, individual rates actually produced vary considerably, much more so for the decelerated and the accelerated versions than for the normal condition. On the average, the fastest condition was produced 61% faster than the normal one, whereas deceleration decreased the rate only by 33%, though individual results were much slower than that, especially for speakers 4 and 7.

Figure 2 shows results of analysis for all tempos produced by female speaker 1 uttering the sentence 1 *Am nächsten Tag fuhr ich nach Husum* - 'On the next day I went to Husum'. The figure displays the following, from the top to the bottom: the speech waveform, the F0 contour (+signs, extracted; solid line, model-based), the German SAMPA phone segmentation, the underlying phrase, and accent commands. The boundaries of underlying syllables are indicated by vertical dotted lines. Vowels carrying lexical stress are marked by bracketing, for instance (E:). Pauses are marked by underscores '_'.

As can be seen, the prosodic structure as reflected by the configuration of underlying accent commands aligned with accented syllables remains intact throughout all conditions, although the amplitudes Aa, durations, and alignments of commands vary. There is a tendency for Aa as well as for the durations of accent commands to become smaller as speed increases. The pause after 'Tag' only disappears in the fastest condition. The declination lines in all utterances can be modeled by a single-phrase command preceding each utterance about 475 ms before the segmental onset. This is slightly smaller than the ideal value of 500 ms. It has to be taken into account that the automatic extraction method aims at pertaining a global optimum of fit for the entire phrase component. This may lead to the phrase command occurring closer to the segmental onset of the phrase.

Figure 3 shows results of analysis by male speaker 4 in the very slow (top) and normal (bottom) conditions. Speaker 4 was one of the subjects who produced his slowest version at around one third the speed of the normal one. Obviously, the sentence was chunked into single words, even syllables, associated with underlying accent commands and often separated by pauses. Comparison with the normal condition by the same speaker shows that low speed does not entail that the accent command amplitude Aa increases, only the number of accent commands rises, as well as the number of phrase commands, as each word virtually becomes a prosodic phrase.

Figure 4 displays another extreme case from the upper limit of the tempo range. It shows sentence 2 uttered by female speaker 6 at fast and very fast speeds. The prosodic structure reflected by the underlying accent commands in the fast version gives way to an almost completely flat F0 contour and a very high value of Fb in the very fast condition. Obviously, the articulation rate becomes so high that proper F0 control can no longer be executed. Overall, we find moderate correlations (Pearson's $r = 0.44$, $P < 0.001$) between Fb and the syllable rate of each utterance indicating that the F0 pattern is raised at higher tempos.

All data were analyzed using R [25] and the R packages *lme4* [26], *languageR* [27,28], and JMP [29]. If not indicated otherwise, data were analyzed using linear mixed effect models. Normality was checked by visual inspection of quantile plots. *Speaker* and *sentence* were

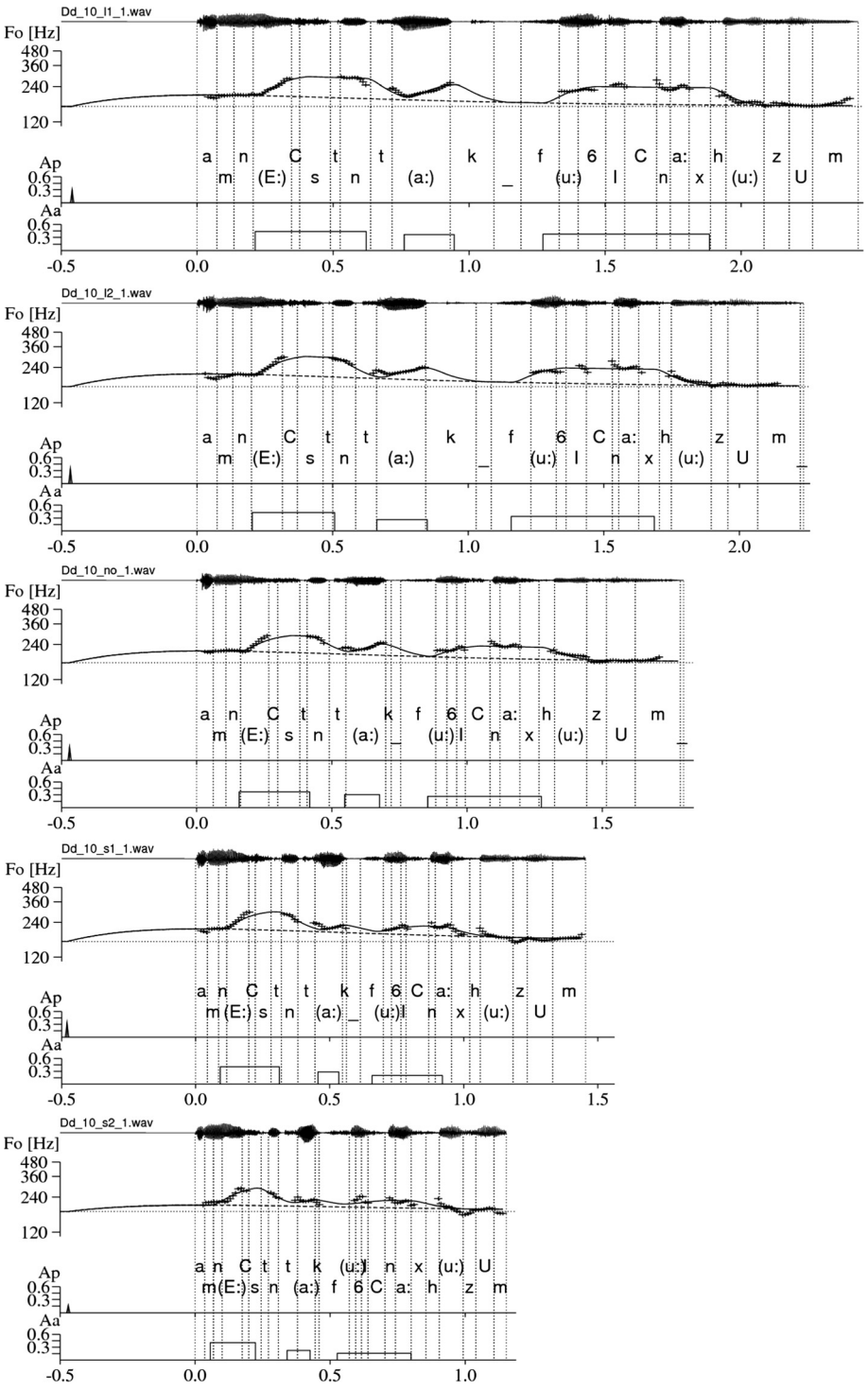

Figure 2 Example of F0 contour decomposition using the Fujisaki model. From the top to the bottom: utterances of sentence 1 *Am nächsten Tag fuhr ich nach Husum* - 'On the next day I went to Husum' produced by female speaker 1 at very slow, slow, normal, fast, and very fast tempo. Lexically stressed vowels are marked by brackets.

treated as random effects, and *intended tempo* as fixed effect. Effects were tested by model comparison between a full model in which the factor in question is entered as either a fixed or a random effect (R code example: lmer (dependent_variable ~ fixed_factor + (1|random_factor1) + (1|random_factor2), data = data)) and a reduced model in

Table 1 Overview of speaker-specific means (*M*) and standard deviations (SD) of syllable rate for the five different intended tempos

Speaker	1 Very slow			2 Slow			3 Normal			4 Fast			5 Very fast		
	M/SD/% normal			M/SD/% normal			M/SD/% normal			M/SD/% normal			M/SD/% normal		
1	4.94	1.86	76	5.57	2.24	85	6.54	2.97	100	7.48	3.33	114	8.98	3.26	137
2	4.72	2.05	67	6.02	2.95	85	7.06	2.89	100	7.29	3.16	103	9.83	3.83	139
3	4.93	1.99	83	5.57	2.40	94	5.91	2.31	100	6.16	2.51	104	9.12	3.74	154
4	2.25	0.83	33	4.63	1.85	67	6.86	2.73	100	7.49	2.84	109	12.04	4.51	176
5	3.98	1.56	85	4.44	1.70	94	4.70	1.60	100	5.61	2.22	119	7.67	2.62	163
6	6.14	2.67	89	5.28	2.55	77	6.87	3.12	100	7.55	3.41	110	11.81	5.39	172
7	1.97	0.53	31	4.58	1.79	71	6.41	2.51	100	7.10	2.64	111	11.88	4.66	185
8	5.66	2.34	82	6.00	2.79	87	6.89	3.01	100	7.30	2.89	106	10.13	4.40	147
9	4.96	1.96	73	6.07	2.74	89	6.83	3.33	100	7.76	3.87	114	10.37	5.00	152
10	3.68	1.68	64	4.73	2.09	82	5.76	2.42	100	8.40	3.43	146	12.01	5.07	209
11	4.70	1.57	71	5.69	1.91	86	6.61	2.53	100	7.51	2.98	114	10.05	4.39	152
12	2.93	1.29	52	3.95	1.86	70	5.63	2.67	100	6.34	2.62	113	8.87	3.70	158
Total	4.24	2.17	67	5.21	2.37	82	6.34	2.78	100	7.16	3.10	113	10.23	4.48	161

Each master column lists mean and standard deviation of syllable rate as well as percentage of in relationship to condition 'normal'.

which the factor in question is excluded (R code example: lmer(dependent_variable ~ 1 + (1|random_factor1) + (1|random_factor2), data = data)). *P* values were retrieved by comparing the results from the two models using ANOVAs (R code: anova(model_full, model_reduced). To assess the relative goodness of fit we indicate Akaike information criterion (AIC) values, which decrease with goodness of fit [30]. *P* values that are considered significant at the $\alpha = 0.05$ level are reported.

4.2 Phrase command parameters

Table 2 summarizes the results obtained from the model comparisons on the Fujisaki phrase level parameters, that is, phrase command magnitude Ap and *phrase duration*, the distance between subsequent phrase commands.

For phrase command magnitude Ap, comparisons between full and reduced models showed a significant difference and both full models exhibited an increased goodness of fit: between-rate variation as well as

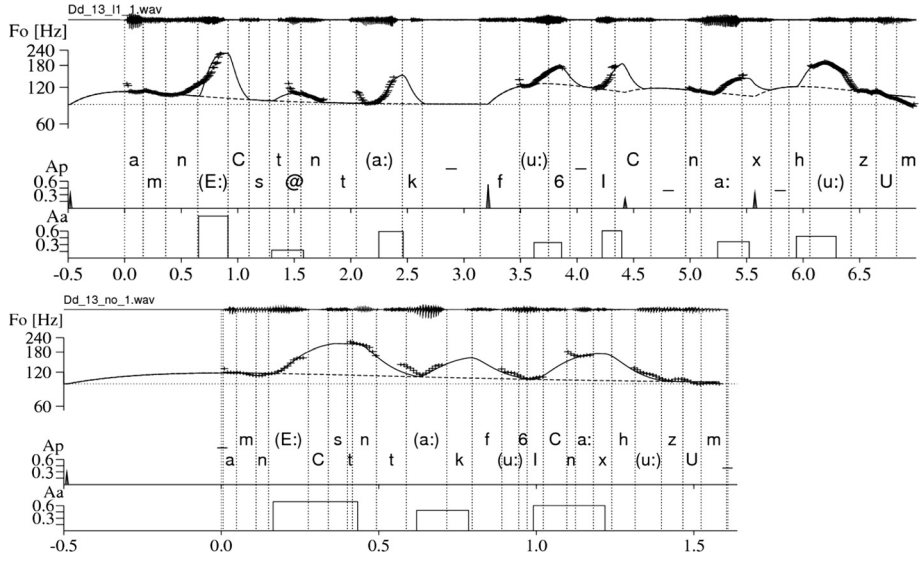

Figure 3 Analysis by male speaker 4 in the very slow and normal conditions. From the top to the bottom: utterances of sentence 1 *Am nächsten Tag fuhr ich nach Husum* - 'On the next day I went to Husum' produced by male speaker 4 at very slow and normal tempo. The time scale of the very slow version was compressed.

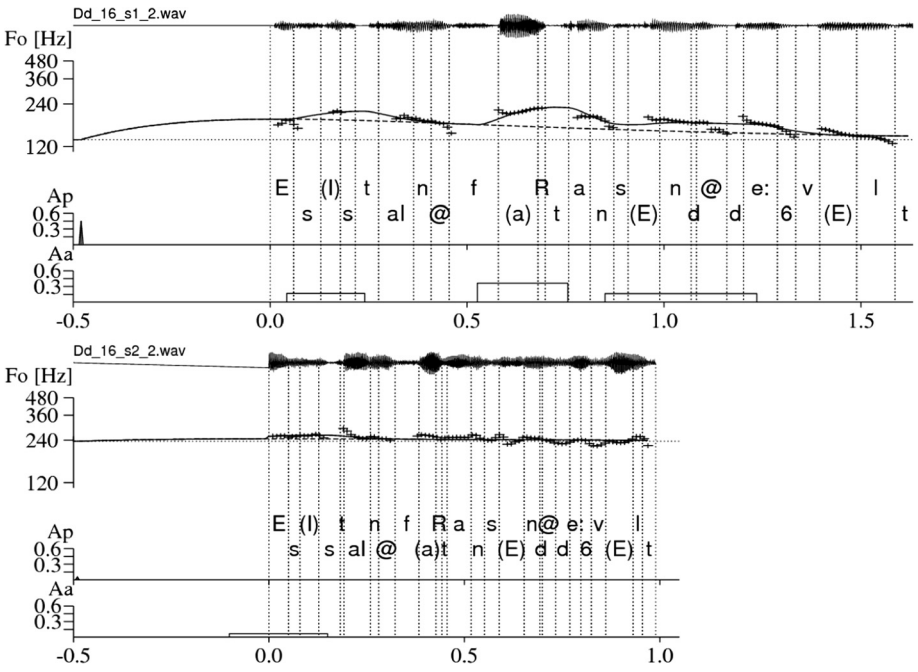

Figure 4 Extreme case from the upper limit of the tempo range. From the top to the bottom: utterances of sentence 2 *Es ist eine Fahrt ans Ende der Welt* - 'It is a journey to the end of the world' produced by speaker 6 at fast and very fast tempo.

between-speaker variation was significant. There was a small but significant interaction between the two factors, i.e., the main effect of intended tempo does not occur independently of speaker. For phrase durations, we found a significant effect of speaker and an interaction of intended tempo and speaker. Figure 5 illustrates values of phrase command magnitude Ap by intended tempo.

Figure 5 reveals that normal speech shows the highest Ap ($M = 0.266$, SD = 0.13), followed by very slow speech ($M = 0.264$, SD = 0.15), and fast intended tempo ($M = 0.263$, SD = 0.12). Very fast speech shows the lowest Ap ($M = 0.17$, SD = 0.12). Visually, if the boxes' notches do not overlap, this can be taken as strong evidence that their medians (solid black lines) differ. Results further revealed a main effect of speaker for phrase duration.

Table 2 Summary of the statistics for the phrase-level Fujisaki model parameters

Fujisaki model parameter	Factor	Result
Phrase command magnitude Ap	Intended tempo	$P < 0.0001$, AIC = −872
	Speaker	$P = 0.0009$, AIC = −872
	Intended tempo*speaker	$P = 0.033$, AIC = −860
Phrase duration	Intended tempo	Not significant
	Speaker	$P = 0.023$, AIC = 1,274
	Intended tempo*speaker	$P = 0.009$, AIC = 1,291

4.3 Accent command parameters

Table 3 summarizes the results obtained from the model comparisons on the Fujisaki accent level parameters, viz. accent command amplitude Aa, the difference between accent command onset time T1 and stressed syllable onset time (*t1relon*), as well as the difference between accent command onset time T1 and vowel onset time of a stressed syllable (*t1relvon*). For the first parameter, only stressed syllables were examined; for the latter two parameters, only stressed syllables that feature an F0 rise, that is, N↑ intonemes, were included in the analyses.

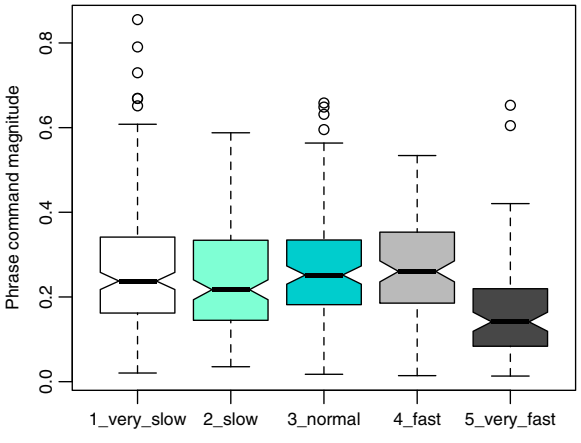

Figure 5 Boxplots of phrase command magnitude by intended tempo. Very slow (white), slow (turquoise), normal (blue), fast (gray), and very fast (dark gray). The white circles denote outliers.

Table 3 Summary of the statistics for the accent-level Fujisaki model parameters

Fujisaki model parameter	Factor	Result
Accent command amplitude Aa	Intended tempo	$P < 0.0001$, AIC = −838
	Speaker	$P < 0.0001$, AIC = −838
	Intended tempo*speaker	$P < 0.0001$, AIC = −894
t1relon (distance between T1 and syllable onset)	Intended tempo	$P < 0.0001$, AIC = −1,282
	Speaker	$P < 0.0001$, AIC = −1,282
	Intended tempo*speaker	$P < 0.0001$, AIC = −1,295
t1relvon (distance between T1 and nuclear vowel onset)	Intended tempo	$P < 0.0001$, AIC = −1,427
	Speaker	$P < 0.0001$, AIC = −1,427
	Intended tempo*speaker	Not significant

Note that the latter two parameters were correlated ($r(496) = 0.73$, $P < 0.0001$; $R^2 = 0.53$, $F(1,494) = 550$, $P < 0.001$); *t1relon* explained 53% of the variability in *t1relvon*.

As for accent command amplitude Aa, the model with intended tempo as a fixed effect provided an improved goodness of fit, which means that between-rate variation was significant. However, descriptive statistics did not reveal a straightforward connection between speech rate and accent command amplitude Aa (very slow $M = 0.33$, SD = 0.15; slow $M = 0.35$, SD = 0.16; normal $M = 0.35$, SD = 0.16; fast $M = 0.35$, SD = 0.16; very fast $M = 0.26$, SD = 0.12). There was a significant effect of speaker. Results further revealed an interaction of *intended tempo*speaker*. Figure 6 shows the box plots of the 12 speakers' accent command amplitude values.

The interaction obtained for intended tempo*speaker becomes evident in Figure 6: whereas speaker 4, for example, exhibited a trend of increasing amplitudes the faster he speaks (except for the very fast condition), speaker 11 performed conversely: the faster his speech, the lower the accent command amplitudes. Given the interaction of *rate*speaker*, the main effects are no longer readily interpretable. To test for the simple effect of intended tempo, we processed 12 ANOVAs, one for each speaker. Only 3 of the 12 ANOVAs showed significant effects of intended tempo (Bonferroni adjusted for speaker, $\alpha = 0.0042$; speakers 2, 4, and 11). Correlation analysis between Aa and speech rate in syllables/s only yielded significant dependencies for four of the speakers with a weak Pearson's $r < -0.3$ indicating a compressed F0 range at higher rates. The syllabic distance between accent commands increases with speed: At the normal rate, subjects produce on average one accent command every 3.1 syllables, at very slow tempo every 2.7 syllables, and at very fast speed every 4.6 syllables.

As for the temporal distance between accent command onset and the onset of a stressed syllable (t1relon), the model with *speaking rate* as a fixed effect provided an improved goodness of, which means that between-rate variation was again significant. Descriptive statistics showed that the faster the speech, the smaller the distance between accent command onset and syllable onset

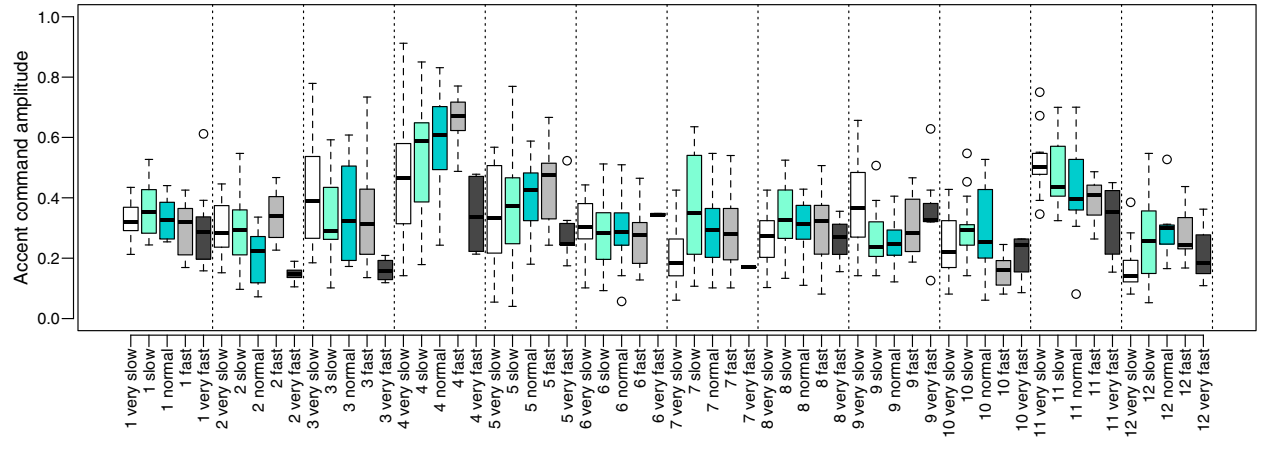

Figure 6 Box plots of the 12 speakers' accent command amplitude values. Very slow (white), slow (turquoise), normal (blue), fast (gray), and very fast (dark gray). The white circles denote outliers. The length of the whiskers (dashed lines) either indicates 1.5 times the interquartile range (IQR) or the minimum/maximum data points in cases where they are closer to the box (indicating the IQR) than 1.5 IQR.

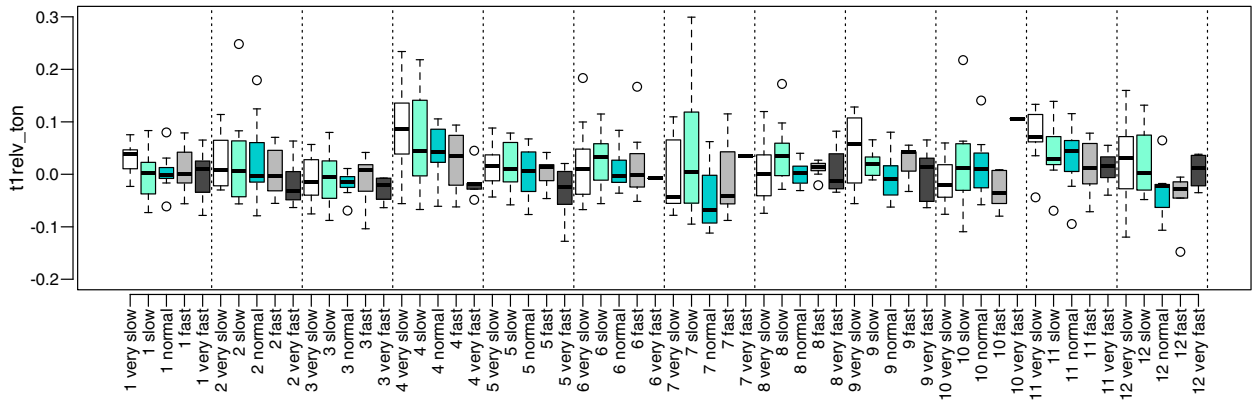

Figure 7 Box plots of the 12 speakers' *t1relvon*. Very slow (white), slow (turquoise), normal (blue), fast (gray), and very fast (dark gray) tempo. The white circles denote outliers. The length of the whiskers (dashed lines) either indicates 1.5 times the interquartile range (IQR) or the minimum/maximum data points in cases where they are closer to the box (indicating the IQR) than 1.5 IQR.

(very slow $M = 0.14$, SD = 0.09; slow $M = 0.12$, SD = 0.08; normal $M = 0.09$, SD = 0.05; fast $M = 0.08$, SD = 0.05; very fast $M = 0.05$, SD = 0.04). In other words, local rises occur earlier in the syllable the faster a person speaks. Results indicated that there was a significant effect of speaker and interaction of intended tempo*speaker. To test for the simple effect of intended tempo, we again processed 12 ANOVAs, one for each speaker. Again, only 3 of the 12 ANOVAs showed significant effects of intended tempo (Bonferroni adjusted for speaker, $\alpha = 0.0042$; speakers 4, 5, and 11).

Concerning the temporal distance between accent command, onset time T1, and the vowel onset within a stressed syllable (*t1relvon*), the model with intended tempo as a fixed effect provided an improved goodness of fit; hence, we interpret that between-rate variation was significant. Overall results showed a similar trend as for t1relon, albeit somewhat less straightforward: the faster the speech, the earlier the local F0 rises relative to the vowel onset (very slow $M = 0.023$, SD = 0.06; slow $M = 0.026$, SD = 0.073; normal $M = 0.003$, SD = 0.05; fast $M = 0.003$, SD = 0.05; very fast $M = -0.006$, SD = 0.04). There was also a significant effect of speaker. Given that intended tempo and speaker do not interact, the main effect of intended tempo occurs independently of the factor speaker. Figure 7 shows the box plots of the 12 speakers' *t1relvon* values.

Figure 7 reveals that most speakers exhibited the trend mentioned above: the faster the speech, the earlier the rise relative to the vowel onset. This is particularly evident for speakers 4, 9, and 11.

Finally, we examined whether the accent command is more strongly anchored to the syllable boundaries or the boundaries of the nuclear vowel. If we calculate correlations between T1 and T2 on the one hand and syllable or vowel onset and offset times on the other hand, Pearson's *r* values are very close to unity since the accent commands are already aligned with the closest syllable. Therefore, we examined means and standard deviations of all timing difference measures, namely

- The difference between accent command onset time T1 and (1) the syllable onset time (*t1relon*), (2) the syllable offset time (*t1reloff*), (3) the vowel onset time (*t1relvon*), and (4) the vowel offset time (*t1relvoff*)
- The difference between accent command onset time T2 and (1) the syllable onset time (*t2relon*), (2) the syllable offset time (*t2reloff*), (3) the vowel onset time (*t2relvon*), and (4) the vowel offset time (*t2relvoff*).

We calculated these measures separately for rising F0 movements (N↑ intonemes) as well as falling F0 movements (I↓ intonemes) at accented syllables with results displayed in Table 4. As can be seen, rising F0 movements occur on average 104 ms after syllable onset and

Table 4 Means (*M*) and standard deviations (SD) in milliseconds for distance measures between accent commands and syllables or nuclear vowels, respectively

Distance measure	Rising F0 (N↑ intonemes)			Falling F0 (I↓ intonemes)		
	N	*M* [ms]	SD [ms]	*N*	*M* [ms]	SD [ms]
t1relon	513	104	76	163	−221	168
t2relon	513	435	166	163	107	88
t1reloff	513	−156	81	163	−462	214
t2reloff	513	175	118	163	−134	118
t1relvon	496	12	60	150	−287	176
t2relvon	496	344	147	150	42	89
t2relvoff	496	226	131	150	−78	96
t1relvoff	496	−106	64	150	−407	199

12 ms after vowel onset. On average, falling F0 movements start 107 ms after syllable onset and 42 ms after vowel onset. These cases are marked in red as they are the ones exhibiting the smallest standard deviations. This suggests that F0 rises and falls are most closely linked to the syllable or vowel onsets. For rising F0 movements, the alignment with the vowel seems to have only a minor advantage, that is, a lower SD. This confirms the observations made earlier in this section regarding marginal differences between *t1relon* and *t1relvon*.

5 Conclusions

The current paper examined the relationship between the F0 contour and speech rate. We employed the Fujisaki model for decomposing F0 contours into utterance level, phrase level, and syllable level components, that is, the base frequency Fb, phrase commands, and accent commands, respectively. We found that only at the extreme ends of the tempo range the prosodic structure disintegrates. Otherwise, the configuration in terms of phrase and accent command numbers and positions remains relatively unchanged, and speech rate has mostly an influence on the amplitudes and exact timings of commands.

In general, we found the following trends: The base frequency Fb increases with speech rate. The phrase duration measured as the distance between consecutive phrase commands decreased with speech rate. Although we expect phrases to contain more syllables at higher speech rate, these syllables are of shorter duration. Unless phrases are merged, phrase duration will therefore also decrease. As the examples in Figure 2 indicated, we do not necessarily see an increase in the numbers of phrase commands at low speed, unless the utterance is broken into very small chunks like in Figure 3, top. This tendency might well be due to the character of the underlying reading material which contains mostly short phrases. The deep boundaries are produced similarly at all tempos, and low speed does not give rise to new phrase commands at shallow boundaries which are rather marked by short pauses.

With respect to accent command amplitude Aa, our outcome suggests that despite a slight trend for Aa to decrease at higher tempos, the influence of the speaker on this parameter by far outweighs that of speech rate. This means that speakers have idiosyncratic ways of manipulating F0 parameters as a function of speech rate. For some speakers, for example, accent command amplitude increases the faster they speak (speaker 4, Figure 6), whereas for others, Aa decreases (speaker 11, Figure 6). We also found that speakers tend to produce fewer accent commands at higher speech rates. This indicates that a high articulatory rate imposes limitations on the frequency of F0 gestures.

As regards the alignment of F0 gestures with the syllable, our results suggest that increased tempo also leads to an early execution of F0 gestures. These observations are similar regardless of whether we anchor the accent command to the syllable onset or the nuclear vowel onset.

It is possible that the reading task underlying our data might have affected the outcome of our analyses, as people repeated the same sentences over and over again. Future work will examine other speech materials produced with more natural tempo variations.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Computer Science and Media, Beuth University Berlin, Luxemburger Str. 10, Berlin 13353, Germany. [2]Phonetic Laboratory, University of Zurich, Rämistrasse 71, Zürich 8006, Switzerland.

Received: 16 January 2014 Accepted: 23 July 2014
Published: 13 August 2014

References

1. F Goldman-Eilser, *Psycholinguistics: Experiments in Spontaneous Speech* (Academic, New York, 1968)
2. ACM Rietveld, CC Gussenhoven, Perceived speech rate and intonation. J. Phonetics **15**, 273–285 (1987)
3. J Caspers, VJ van Heuven, Effects of time pressure on the phonetic realization of the Dutch accent-lending pitch rise and fall. Phonetica **50**, 161–171 (1993)
4. DR Ladd, D Faulkner, H Faulkner, A Schepman, Constant segmental anchoring of *F0* movements under changes in speech rate. J. Acoust. Soc. Am. **106**, 1543–1554 (1999)
5. H Fujisaki, K Hirose, Analysis of voice fundamental frequency contours for declarative sentences of Japanese. J. Acoust. Soc. Japan (E) **5**(4), 233–241 (1984)
6. H Fujisaki, K Hirose, *Modeling the dynamic characteristics of voice fundamental frequency with applications to analysis and synthesis of intonation: preprints of the working group on intonation* (13th International Congress of Linguistics, Tokyo, 1982), pp. 57–70
7. H Mixdorff, *Intonation patterns of German—model-based quantitative analysis and synthesis of F0-contours* (D.Eng, Thesis, TU Dresden, 1998)
8. A Leemann, *Swiss German Intonation Patterns* (Benjamins, Amsterdam/New York, 2012)
9. DP Kuehn, KL Moll, A cineradiographic study of VC and CV articulatory velocities. J. Phonetics **23**(4), 303–320 (1976)
10. B Lindblom, Explaining phonetic variation: a sketch of the H&H theory, in *Speech production and speech modelling*, ed. by WJ Hardcastle, A Marchal (Kluwer, 1990), pp. 403–439
11. JE Flege, Effects of speaking rate on tongue position and velocity of movement in vowel production. JASA **84**, 901–916 (1988)
12. AV Isačenko, HJ Schädlich, *Untersuchungen über die deutsche Satzintonation* (Akademie-Verlag, Berlin, 1964)
13. E Stock, C Zacharias, *Deutsche Satzintonation* (VEB Verlag Enzyklopädie, Leipzig, 1982)
14. H Mixdorff, C Widera, *Perceived prominence in terms of a linguistically motivated quantitative intonation model* (Proc. Eurospeech 2001, Aalborg, Denmark, 2001), pp. 403–406
15. H Mixdorff, O Jokisch, *Building an integrated prosodic model of German, vol. 2* (Proceedings of Eurospeech 2001, Aalborg, Denmark, 2001), pp. 947–950
16. H Mixdorff, H Fujisaki, Production and perception of statement, question and non-terminal intonation in German. Proc. ICPhS, Stockholm **2**, 410–413 (1995)
17. N Amir, H Mixdorff, O Amir, D Rochman, GM Diamond, T Isserles, S Abramson, HR Pfitzinger, *Unresolved anger: prosodic analysis and classification of speech from a therapeutical setting* (Proceedings of Speech Prosody 2010, Chicago, USA, 2010)

18. V Dellwo, P Wagner, Relationships between speech rate and rhythm, in *Proceedings of the ICPhS 2003* (Barcelona, 2003)
19. V Dellwo, I Steiner, B Aschenberner, J Dankovicova, P Wagner, The BonnTempo-corpus & BonnTempo-tools: a database for the study of speech rhythm and rate, in *Proceedings of ICSLP 2005*, 2005
20. B Schlink, *Selbs Betrug* (Diogenes Verlag, Zurich, 1994)
21. HR Pfitzinger, Local speech rate perception in German speech. Proc. ICPhS **1999**, 893–896 (1999)
22. P Boersma, Praat, a system for doing phonetics by computer. Glot Int. **5**, 341–345 (2001)
23. H Mixdorff, *A novel approach to the fully automatic extraction of Fujisaki model parameters, vol 3* (Proceedings of ICASSP 2000, Istanbul Turkey, 2000), pp. 1281–1284
24. H Mixdorff, *FujiParaEditor*, 2009. http://public.beuth-hochschule.de/~mixdorff/thesis/fujisaki.html
25. R Core Team, *R, A language and environment for statistical computing, R foundation for statistical computing*, 2013. *version 3.0.0*, http://www.R-project.org
26. DM Bates, M Maechler, *lme4: linear mixed-effects models using S4 classes*, 2009. R package version 0.999375-32
27. RH Baayen, *Analyzing linguistic data: a practical introduction to statistics using R* (CUP, Cambridge, 2008)
28. RH Baayen, *LanguageR: data sets and functions with analyzing linguistic data: a practical introduction to statistics using R*, 2009. R package version 0.955
29. JMP, *Version 9.0* (SAS Institute Inc, Cary NY, 1989–2007)
30. R Kliegl, P Wei, M Dambacher, M Yan, X Zhou, Experimental effects and individual differences in linear mixed models: estimating the relationship between spatial, object, and attraction effects in visual attention. Front. Psychol. **1**(238), 1–12 (2011)

doi:10.1186/s13636-014-0033-6
Cite this article as: Mixdorff *et al.*: The influence of speech rate on Fujisaki model parameters. *EURASIP Journal on Audio, Speech, and Music Processing* 2014 **2014**:33.